УДК 517.958+519.242

ОБ ОЦЕНИВАНИИ ДИФФЕРЕНЦИАЛЬНОЙ ЭНТРОПИИ СЛУЧАЙНЫХ ВЕКТОРОВ

ASSESSMENT OF THE DIFFERENTIAL ENTROPY OF RANDOM VECTORS

Геворгян Гарник Гургенович, аспирант каф. «Прикладная математика»,Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, Россия, 620002, г. Екатеринбург, ул. Мира, 19. E-mail: garnik.ggg@gmail.com. Тел.: +7(912)204-94-75
Тырсин Александр Николаевич, д-р. техн. наук, заведующий каф. «Прикладная математика»,Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, Россия, 620002, г. Екатеринбург, ул. Мира, 19. E-mail: at2001@yandex.ru. Тел.: +7(922) 100-74-52

Garnik G. Gevorgyan, PhD student, Department Chairman «Applied mathematics», Ural Federal University named after the first President of Russia B.N.Yeltsin, 620002, Mira street, 19, Ekaterinburg, Russia. E-mail: garnik.ggg@gmail.com. Ph.: +7(912)204-94-75
Alexander N. Tyrsin, Doctor Sc., DepartmentChairman «Applied mathematics», Ural Federal University named after the first President of Russia B.N.Yeltsin,620002, Mira str., 19, Ekaterinburg, Russia. E-mail: at2001@yandex.ru. Ph.: +7(922) 100-74-52

Аннотация: Рассмотрены особенности оценивания дифференциальной реализации случайных векторов для практического использования. Исследован вопрос устойчивости процедуры оценивания к присутствию в выборках аномальных наблюдений. Рассмотрена возможность использования дифференциальной энтропии для выборок случайных векторов, некоторые компоненты которых представлены в сгруппированном виде как дискретные величины. Приведеныпримерыоцениваниядифференциальнойэнтропииспомощьюпредложенныхалгоритмов.
Abstract: Hereby, the features of assessment of differential realization of random vectors for practical use are considered. The question of the stability of the assessment procedure for the presence of anomalous observations in the samples was investigated. The possibility of using differential entropy for samples of random vectors is considered, some components of which are presented in grouped form as discrete quantities. Examples are given of differential entropy assessment with the help of the proposed algorithms.
***Ключевые слова*:** дифференциальная энтропия; выборка; оценка; аномальное наблюдение; закон распределения; дискретная случайная величина; метод Монте-Карло.
***Keywords*:** differential entropy; sample; assessment; anomalous observation; distribution law; discrete random variable; Monte Carlo method.

ВВЕДЕНИЕ

В [1] была введена дифференциальная энтропия многомерной случайной величины $Y = (Y_1, Y_2, .., Y_m)$ с плотностью распределения $p_Y(x_1, \dots, x_m)$, равная

$$H(Y) = -\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p_Y(x) \ln p_Y(x) dx. \quad (1)$$

В частном случае для одномерной непрерывной случайной величины X дифференциальная энтропия определяется по формуле

$$H(X) = -\int_{-\infty}^{+\infty} p_X(x) \ln p_X(x) dx, \quad (2)$$

где $p_X(x)$ – плотность распределения случайной величины X.

ОЦЕНИВАНИЕ ЭНТРОПИЙНОЙ МОДЕЛИ

На основе (1) в [2] предложенная энтропийная модель описания многомерных стохастических систем, представляемых в виде случайных векторов. Для случайного вектора Y с произвольным распределением получена формула [2]

$$H(\mathbf{Y}) = \sum_{i=1}^{m} \ln \sigma_{Y_i} + \sum_{i=1}^{m} \kappa_i + \frac{1}{2} \sum_{k=2}^{m} \ln(1 - R^2_{Y_k/Y_1 Y_2 \dots Y_{k-1}}), \quad (3)$$

где $\kappa_i = H\left(\frac{Y_i}{\sigma_{Y_i}}\right)$ – энтропийный показатель типа закона распределения случайной величины Yi, $i = 1, 2, \dots, m$; $R^2_{Y_k/Y_1 Y_2 \dots Y_{k-1}}$ – индексы детерминации регрессионных зависимостей между компонентами случайного вектора Y, $k = 2, 3, \dots, m$.

Достоинством формулы (3) является то, что энтропийное моделирование многомерных стохастических систем на ее основе не требует знания или определения закона распределения многомерной случайной величины Y, что практически нереализуемо в реальных задачах. При этом в отличие от методов многомерного статистического анализа, здесь не теряется формальная строгость и соответствие модели (3) реальным экспериментальным данным. Это позволяет использовать формулу (3) для моделирования и исследования реальных многомерных стохастических систем и процессов по экспериментальным данным ограниченного объема.

Согласно (3) параметрами энтропийной модели являются: средние квадратические отклонения σ_{Y_i} компонент Y_i; энтропийные показатели κi законов распределений, $i = 1, 2, \ldots, m$; индексы детерминации $R^2_{Y_k/Y_1Y_2\ldots Y_{k-1}}$ регрессионных зависимостей между компонентами случайного вектора Y, $k = 2, 3, \ldots, m$.

Оценивание индексов детерминации $R^2_{Y_k/Y_1Y_2\ldots Y_{k-1}}$ описано в [3]. Рассмотрим вопрос оценки энтропийного показателя по формуле (3), где $\sigma_X = 1$. В настоящее время предложен ряд алгоритмов для решения данной задачи. Они, как правило, используют априорную информацию о свойствах случайной величины X. Исследуем алгоритм [4], не требующий априорных сведений о случайной величине X. Он основан на формуле

$$\hat{\kappa} = \hat{H}(X) = -\sum_{j=1}^{L} \hat{p}_j \ln \hat{p}_j + \ln h, \qquad (4)$$

где $\hat{p}_j = \frac{n_j}{n}$ – оценка вероятности попадания в j-й интервал; L – число интервалов разбиения равной длины h, непрерывно покрывающих диапазон данных x_i, $i = 1, 2, \ldots, n$; n_j – число наблюдений, попавших в j-й интервал; n – объем выборки.

Слагаемое lnh в (4) введено для устранения смещения формулы информационной энтропии. Основная проблема использования формулы (4) – выбор количества интервалов *L*, которое зависит от типа распределения случайной величины *X*. При слишком малом числе интервалов *L* оценка $\hat{H}(X)$ будет завышенной, а при слишком большом числе интервалов – заниженной.

Методом статистических испытаний Монте-Карло [4] для тридцати типов распределений была получена эмпирическая формула для оптимального количества интервалов

$$\hat{L} = 1{,}072n^{0{,}968-0{,}231I_{0,8}} - 2{,}098\gamma_1 - 1{,}789, \quad (5)$$

где γ_1 – коэффициент асимметрии, $I_{0,8} = x_{0,9} - x_{0,1}$ – интердецильный размах, $x_{0,1}$ и $x_{0,9}$ – квантили эмпирического распределения уровней 0,1 и 0,9.

Формула (5) достаточно точно описывает зависимость количества интервалов от объема выборки, коэффициента асимметрии и интердецильного размаха. Некоторые результаты статистических испытаний (число испытаний равно 50000) приведены в табл. 1. Здесь приведены 95%-е доверительные интервалы оценок энтропийных показателей по формулам (4), (5).

Исследования показали, что оценки неустойчивы к выбросам. В табл. 2 приведены 95%-е доверительные интервалы оценок энтропийных показателей для нескольких распределений по выборкам, в которых три наибольших значения были увеличены в два раза. Видим, результаты оценивания значительно смещены относительно теоретических значений κ. Можно отметить, что с увеличением объема выборки *n* смещение уменьшается. Аналогичные результаты получены практически для всех распределений.

Таблица 1.

95%-е доверительные интервалы оценок $\hat{\kappa}$энтропийных показателей κ по формулам (4), (5)

Распределение $F_X(x)$	κ	$\hat{\kappa}$		
		n=250	n = 500	n=1000
Вейбулла	1	(1,000; 1,026)	(0,990; 1,008)	(0,989; 1,003)
Гамбела	1,329	(1,349; 1,365)	(1,337; 1,347)	(1,332; 1,338)
Гамма-распределение	1	(1,000; 1,026)	(0,990; 1,009)	(0,989; 1,002)
Лапласа	1,347	(1,349; 1,365)	(1,346; 1,358)	(1,346; 1,353)
Логистическое	1,405	(1,402; 1,411)	(1,403; 1,409)	(1,404; 1,408)
Логнормальное	0,649	(0,639; 0,717)	(0,632; 0,693)	(0,633; 0,679)
Нормальное	1,419	(1,415; 1,423)	(1,418; 1,422)	(1,419; 1,421)
Парето	0,779	(0,783; 0,836)	(0,762; 0,805)	(0,763; 0,795)
Равномерное	1,24	(1,224; 1,235)	(1,230; 1,238)	(1,234; 1,239)
Тригонометрическое	1,395	(1,388; 1,395)	(1,389; 1,394)	(1,391; 1,394)
Экспоненциальное	1	(1,001; 1,027)	(0,990; 1,009)	(0,990; 1,003)

Для обеспечения устойчивости оценивания энтропийных показателей можно воспользоваться известными процедурами цензурирования и винзорирования [6] исходной выборки x_i, $i = 1, 2, \ldots, n$. Исследования показали устойчивость оценок после устранения выбросов.

Таблица 2.

95%-е доверительные интервалы оценок $\hat{\kappa}$энтропийных показателей κ по формулам (4), (5)

Распределение $F_X(x)$	κ	$\hat{\kappa}$		
		n = 250	n = 500	n = 1000
Логистическое	1,405	(1,293; 1,313)	(1,329; 1,341)	(1,356; 1,364)
Нормальное	1,419	(1,347; 1,361)	(1,374; 1,382)	(1,394; 1,398)
Тригонометрическое	1,395	(1,188; 1,200)	(1,287; 1,293)	(1,344; 1,348)

Еще одной проблемой энтропийной модели (4) является невозможность ее использования для дискретных случайных величин. Для расчета энтропийных показателей необходимо доопределить дискретные случайные величины до непрерывных. Это неоднозначная процедура.

Рассмотрим произвольную дискретную случайную величину X, которая задана рядом распределения (табл. 3).

Таблица 3.

Ряд распределения дискретную случайную величину X

X	x_1	x_2	...	x_M
$P(X = x_k)$	p_1	p_2	...	p_M

Задача состоит в том, чтобы от дискретной случайной величины X перейти к непрерывной случайной величине Z. Рассмотрим ситуацию, когда фактические значения случайной величины занимают непрерывную область, но с помощью, как правило, экспертных процедур задают некоторое фиксированное множество значений $x_1, x_2, \ldots, x_M$.

В данной ситуации необходимо сделать некоторое разумное предположение о формировании значений случайной величины X. В случае интервального группирования ошибки округления для каждого значения x_k можно рассматривать как равномерно распределенные непрерывные случайные величины. Тогда искомая случайная величина Z будет иметь кусочно-постоянную плотность вероятности $p_Z(x)$, равную

$$p_Z(x) = \begin{cases} 0, & -\infty < x \le x_{1,0}, \\ \frac{p_1}{x_2 - x_1}, & x_{0,1} < x \le x_{1,2}, \\ \frac{p_k}{x_k - x_{k-1}}, & x_{k-1,k} < x \le x_k, \\ \frac{p_k}{x_{k+1} - x_k}, & x_k < x \le x_{k,k+1}, \\ \frac{p_M}{x_M - x_{M-1}}, & x_{M-1,M} < x \le x_{M,M+1}, \\ 0, & x_{M,M+1} < x < +\infty, \end{cases} \quad (6)$$

где $x_{0,1} = x_1 - \frac{x_2 - x_1}{2}$, $x_{M,M+1} = x_M + \frac{x_M - x_{M-1}}{2}$, $x_{j,j+1} = \frac{x_j + x_{j+1}}{2}$, $j = 1, 2, \ldots, M-1$, $k = 2, 3, \ldots, M-1$.

Очевидно, что непосредственная подстановка выражения (7) в (3) позволяет определить $H(Z)$.

ВЫВОДЫ

1. Установлена неустойчивость оценивания энтропийного показателя к присутствию в выборках аномальных наблюдений. Для устранения этого недостатка предложено воспользоваться процедурами цензурирования и винзорирования [6] исходной выборки.

2. Показана возможность распространения энтропийного подхода на основе модели (3) для выборок случайных векторов, некоторые компоненты которых представлены в сгруппированном виде как дискретные величины.

Исследование выполнено при поддержке РФФИ, грант № 17-01-00315а.

БИБЛИОГРАФИЧЕСКИЙСПИСОК

1. Shannon C.E. A Mathematical Theory of Communication. The Bell System Technical Journal. 1948. V. 27. P. 379-423, 623-656.
2. Тырсин А.Н. Энтропийное моделирование многомерных стохастических систем. Воронеж: Научная книга, 2016. 156 с.
3. Тырсин А.Н., Калев О.Ф., Яшин Д.А., Лебедева О.В. Оценка состояния здоровья популяции на основе энтропийного моделирования. Математическая биология и биоинформатика. 2015. Т. 10. Вып 1. С. 206-219. doi: 10.17537/2015.10.206.
4. Тырсин А.Н., Клявин И.А. Повышение точности оценки энтропии случайных экспериментальных данных // Системы управления и информационные технологии. 2010. № 1(39). С. 87-90.
5. Михайлов Г.А., Войтишек А.В. Численное статистическое моделирование. Методы Монте-Карло. – М.: Издательский центр «Академия», 2006. 368 с.
6. Смоляк С.А., Титаренко Б.П. Устойчивые методы оценивания. М.: Статистика, 1980. 208 с.